MSDS600 Week 4 Assignment - Nathan Worsham

The six data sets given to analyze all have one thing in common: that the Mean and Median are very close if not the same on all distributions. If there is an exception, it would be N1 and N2 whose Mean and Median are not exactly right on top of each other but still relatively very close.

Binomial.csv

statistic	value
Mean	70.17
Median	70
Mode	68
Standard Deviation	4.689325
Minimum	57
Maximum	84
Range	27
1st Quartile	67
3rd Quartile	73
IQR	6
# of Values	1000

ln.csv

statistic	value
Mean	18.99
Median	19
Mode	18
Standard Deviation	4.362612
Minimum	3
Maximum	43
Range	40
1st Quartile	16
3rd Quartile	22
IQR	6
# of Values	1048576

BN1.csv vs BN2.csv

To compute the mean, I took each value multiplied by its number of occurrences:

```
16*-5 + 121*-4 + 182*-3 + 254*-2 + 417*-1 + 307*1 + 581*2 + 397*3
+ 93*4 + 2*5
```

becomes:

```
-80 + -484 + -546 + -508 + -417 + 307 + 1162 + 1191 + 372 + 10 = 1007
```

Then summed the number of occurrences:

statistic	BN1	BN2
Mean	9.994	10.997
Median	9.993	10.998
Mode	10.07391	10.94372
Standard Deviation	2.000357	0.9994714
Minimum	1.781	6.638
Maximum	18.612	15.161
Range	16.83086	8.523442
1st Quartile	8.643	10.321
3rd Quartile	11.343	11.667
IQR	2.700128	1.345798
# of Values	100000	100000

```
16 + 121 + 182 + 254 + 417 + 307 + 581 + 397 + 93 + 2 = 2370
```

And finally divided the sum by the count:

```
1007/2370 = 0.4248945147679325
```

I was able to run this into R by using the following command:

```
sentiment <- c(rep(-5,16),rep(-4,121),rep(-3,182),rep(-2,254),
rep(-1,417),rep(1,307),rep(2,581),rep(3,397),rep(4,93),rep(5,2))
```

Which I could then run the summary function to confirm my results

```
 summary(sentiment)
  Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-5.000000 -1.000000  1.000000  0.424895  2.000000  5.000000
```

So the Mean is slightly in the positive direction. It would seem that this would indicate that during the timeframe I recorded the twitter feed that the general sentiment/mood/attitude of tweets made were slightly more of a positive sentiment than of a negative sentiment. This does seem to agree with the values I received as there was a total of 1380 positive comments versus 990 negative comments. Calculating the Mean makes the assumption that the values of the grouping of words is quantitative, this is problematic as mentioned earlier that the grouping of the words is subjective however it is not without meaning as it does seem to help describe the data set. What I also found interesting is that the Mode is 2 and Median is 1, which the average is below the Median making this a left-skewed distribution data set. So visually we can see that the data set distribution converges on the right, which can also be seen by providing the Mean, Median, and Mode together: pic

If different words were used in the AFINN.txt file